4

Trends in Bioinformatics Technology

Atsushi Nogi and Shotaro Kohtsuki
Affiliated Fellows

4.1 Introduction

Through international cooperation, the Human Genome Project has proceeded at a greater-than-expected speed. Accomplishing the draft sequencing by June 2000, the project has now entered its final stage toward complete sequence determination. Japan's contribution to the project was 6–7%, which corresponded to the budget scale. However, problems have remained, such as the lack of strategy, which has forced Japan into a late start in this area. Nevertheless, research projects such as complete sequencing of mouse full-length cDNA (See Footnote) and rice genome sequencing led by The Institute of Physical and Chemical Research (RIKEN), accomplished in August 2002, and ascidian genome sequencing conducted as a joint research project between Japan and the U.S., completed in December 2002, have driven genome science into a new stage. Now that human genome and mouse cDNA have been sequenced, the focus should be shifted from sequence analyses to comprehensive, systematic functional analyses. Japan has advantages in fundamental technologies such as rice genome and full-length cDNA libraries and, therefore, has great potential to take the leading role both in academic and applied areas.

In recent years, the speed-up and automation of analyzers have resulted in the generation of vast amounts of experimental data, which has imparted a growing importance to information technology. In addition to the improvement of data processing efficiency, knowledge of mathematics and informatics is required to cope with the complicated, comprehensive analyses performed on genome information. In order to work on such issues systematically, a new academic area, bioinformatics, has been created by the interface of bioscience and informatics.

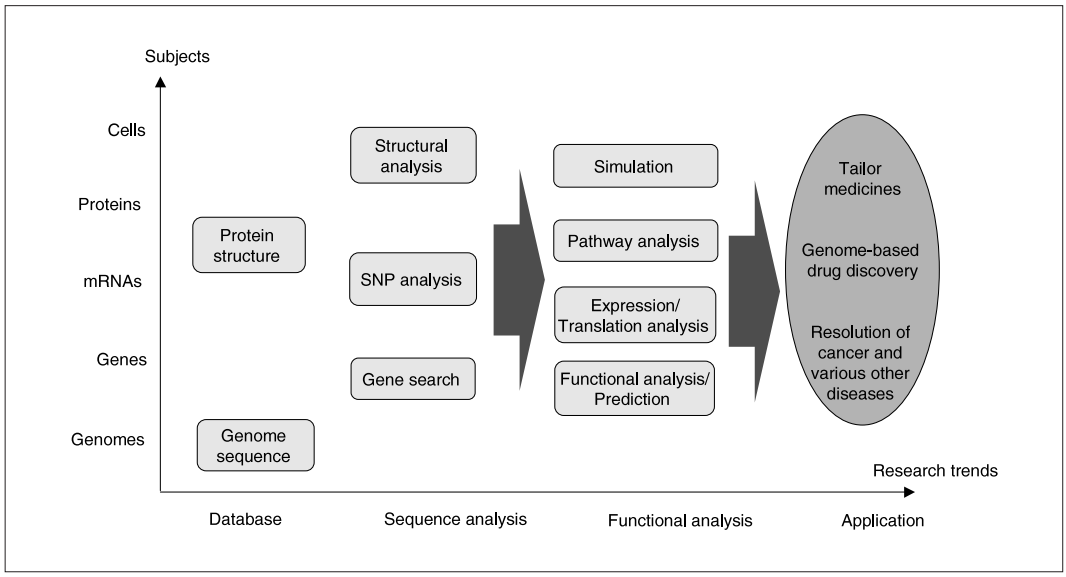
Bioinformatics has been covered in the article of this journal's November 2002 issue, which introduced an overview of the trends in bioinformatics from a bioscience point of view. This report article is a sequel to this, discussing the methodology for the systematic understanding of genome and life phenomena and the up-to-date issue of human resource development.

Footnote:
full-length cDNA: cDNAs are DNAs obtained by excluding unnecessary sequences from genomic DNAs. cDNAs are generated by using mRNAs (messenger RNAs, the genetic information-carrying substances that only contain protein-encoding sequences) as templates. Unlike partial cDNA fragments, full-length cDNAs possess all the information necessary for protein synthesis and is therefore capable of synthesizing proteins. Efficient synthesis of full-length cDNA requires very high skill, in which Japan takes the initiative among other countries.

4.2 Informatics-based approaches in bioinformatics

If we view the issues in genome research from an informatics point of view and attempt their formulation, we soon realize that most of such issues are far beyond the capabilities of existing calculators. This can be attributed to the extremely long lengths of genomic sequences and to the combinatorial nature of the issues. Therefore, such issues must be analyzed by approximate or heuristic means. The development of practical algorithms is the main research area in bioinformatics, and its achievements have greatly

Figure 1: Research trends in bioinformatics



contributed to the success of genomic sequencing. Whole genome shotgun sequencing, an approach employed by Celera Genomics (U.S.), involves the assemblage of tenmillions of random sequence fragments as in a jigsaw puzzle. At first, the approach itself was considered to have great uncertainty, but using top-performance calculators and a unique algorithm, Celera Genomics accomplished sequence determination at an astoundingly high speed, which has strongly demonstrated the power and effectiveness of informatics.

Using approximate means does not necessarily come up with the optimal solution. Therefore, the analysis results require biological evaluation, based on which the algorithms or parameters must be modified. Bioinformatics can be characterized as a tool for narrowing down the vast search domain, which otherwise requires experimental confirmation. It also enables systematic, comprehensive analyses to elucidate the general view of life phenomena, which could not have been seen by analyzing the individual genes.

To date, the main research subjects in bioinformatics were the construction of genomic databases, and the development of data analysis tools and their application techniques. From now on, a higher data-processing capability will be employed in genome research for discovering genes and predicting their functions. As shown in Figure 1, genome research, which started from genome sequencing, has gradually shifted its research subject from DNA to genes, genes to proteins and proteins to individuals, increasing the complexity and diversity of the subjects. The subjects of analyses have shifted from sequences to functions or actions. Research and development in bioinformatics must be in concert with such trends in genome research.

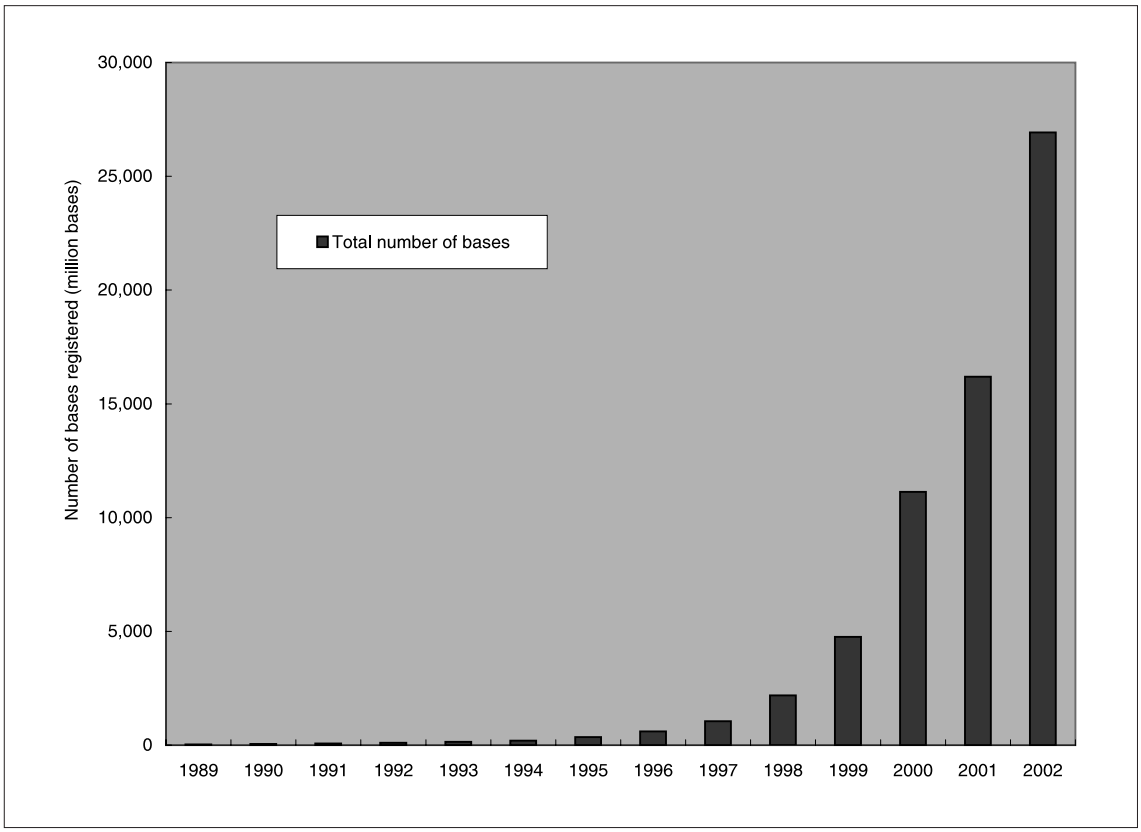
4.2.1 Databases

The majority of the vast amounts of data generated from genome research are registered in public databases, which are available to the public through the Internet. Such data can be viewed as a research foundation, which serves as a starting point for every informatics research. The number of DNA sequence data registered in public genomic databases (GenBank (U.S.), EMBL (Europe), and DDBJ (Japan)) are growing exponentially (Figure 2). Conventionally, databases were used to accumulate primary data such as DNA sequences and protein structural data, but along with the progress of genome projects, they have diversified, incorporating new types of data such as data on single nucleotide polymorphisms (SNP) or mutations and data on gene or protein interactions. In order to perform analyses employing such diverse types of data, further advances in databases and information technology are required.

(1) Annotation of databases

A series of processes for discovering new

Figure 2: Transition in the number of bases registered



Source: Website of DNA Data Bank of Japan

knowledge from various data is called data mining, an important technique in bioinformatics. However, after the experimental data are registered in databases, they are left unprocessed, obstructing the effective application of data mining techniques. To solve this problem, information on their interpretation must be imparted to such data. This process is called annotation. Although computerization of annotation is currently under way, computers alone cannot figure out the optimum annotation, requiring human assistance for its validation. A conference on the annotation of mouse full-length cDNA sequence determined at RIKEN was held in August 2000 at RIKEN in Tsukuba. The annotation process involved 2 weeks of heated debates by participating researchers from various countries. Beyond interpretations of individuals, annotation requires the cooperation of researchers from various organizations for its validation.

(2) Integration of ideas and vocabularies

Each gene or protein is given a customary name during its research process. As a result, the same protein may have been given several different names in different research areas, which require integration. As genome sequencing has been successively accomplished in human and other model organisms, more and more studies are beginning to focus on the entire living world. Conventionally, databases have been constructed independently for each model organism (E. coli, yeast, mouse, etc.), but the lack of criteria for the integration of terms or molecular names have hindered their efficient use.

In order to solve these problems, organization of non-standardized vocabularies has been initiated to enable their systematic description. Such process is called ontology, a process for imparting consistent terms and definitions to ideas. The establishment of ontology has been promoted for each research area, such as gene ontology covering genes, interaction ontology covering molecular and cellular interactions, and signal ontology covering signal transductions. Progress in such integration and systematic classification of vocabularies should increase mutual application of databases, thereby enabling mutual reference among databases of different species.

4.2.2 Homology analysis

Homology analysis is a powerful approach for studying the functions of genes or proteins based on interspecific amino acid sequence similarities. Proteins show higher structural and functional similarities between closely related species. Consequently, if a functionally identified gene with a similar sequence to the target gene can be found in another species, it would serve as a key to predict the functional characteristics of the target gene.

In a homology analysis involving an extremely long sequence such as a genomic sequence, speed is required in finding homologous sequence regions. Typical analysis programs for this purpose are FASTA (Fast Alignment) and BLAST (Basic Local Alignment Search Tool). Comparing these two, BLAST is more commonly used, due to its higher processing speed, while FASTA is used in detailed analyses due to its high detection sensitivity.

Even BLAST or FASTA requires a long time to search through large-scale genome databases. In such cases, parallelization of the searching process is effective. Examples of parallelization methods are the multiprocessor method SMP (Symmetric MultiProcessor), PC clusters and grids. The multiprocessor method involves the loading of multiple CPUs onto a single computer system to increase the processing performance. PC clustering realizes a parallel calculator at low cost by connecting multiple standard PCs through networks. The widely used BLAST program provided by NCBI (National Center for Biotechnology Information: an integral database for biological data in the U.S.) is designed for multiprocessor type parallel calculation, but the multiprocessor method requires expensive hardware costs. Therefore, BLAST designed for the less expensive PC clustering is available for practice.

Another alternative is the sharing of the calculation environment, which is represented by grid. Grid is one of the information technologies receiving great attention in recent years. The BioGrid Project by Osaka University and OBIGrid (discussed later) of Initiative for Parallel Bioinformatics (IPAB) are examples of domestic bio-related grids. OBIGrid provides an environment for using the latest databases and applications required for bioinformatics. By using such programs, individual laboratories no longer need to purchase expensive calculation facilities. Our country must work on the further enrichment of such domestic efforts.

4.2.3 Protein structure analysis

Proteins are functional molecules fundamental to life activity, which are the final products synthesized *in vivo* based on the genomic information. One of the largest objectives in bioinformatics is to understand the relationship between the amino acid sequence and three-dimensional structures of proteins. In these few years, the amount of data concerning protein structure has increased drastically, showing a 10-fold increase from a decade ago.

The three-dimensional structure of a protein is uniquely determined by its amino acid sequence, which means that theoretically the structure of a protein can be predicted from its sequence. After a protein is synthesized in a cell, it is folded into the most energetically stable structure within a few milliseconds to seconds. This process is called folding. The precision of protein structure prediction based on molecular dynamics is still poor, due to the enormous amount of calculation required for this approach.

An example of a practical approach for structure prediction is homology modeling. Homology modeling takes advantage of the fact that proteins with similar sequences also have similar structures. When the target protein has 30% or higher homology with a known protein at a sequence level, the structure of the target protein can be predicted by partially altering the structure of its homologous protein.

Since the prediction of the protein structure has already been realized, the focus of protein research has now shifted into functional analysis. Functional analysis for predicting the functional characteristics of proteins is an important step for drug discovery, and the role of bioinformatics should be extremely important in this area.

4.2.4 Gene Network

The determination of complete genome sequences or protein structures does not mean that we have fully understood the life itself. Our next step is to figure out how genes interact with

each other. To elucidate such network of genes, the vast amount of information that has been accumulated in bioscience needs to be systematized in terms of interaction, so that it can be handled with calculators. Therefore, this area is being focused on as a new application area of bioinformatics.

In genome functional analyses, the whole set of genes expressed in a cell is called transcriptome, and the whole set of proteins synthesized from genes is called proteome. Identifying when and how a gene is expressed would provide an important clue to the identification of its function.

Data obtained from proteome analyses are also useful for other researchers. Swiss Institute of Bioinformatics is promoting database construction of electrophoresis gel images representing the analysis results obtained for proteins contained in experimental samples. However, since proteome data are patentable and directly lead to industrial applications such as drug discovery, their public disclosure is moving towards restriction. Therefore, it is urgent that domestic proteome-related databases are established.

In addition, studying the cell as a dynamic system constructed by genes encoded in the genome is becoming a main research theme in bioinformatics. Kyoto University's KEGG (Kyoto Encyclopedia of Genes and Genomes) system discloses the results from their gene network research in the form of a database. The conventional methodologies for biological research based on the description of gene functions are inadequate for gene network studies, and must be integrated with informatics.

4.3 Issues in bioinformatics from the viewpoint of information technology

Most analysis tools (software) used in bioinformatics are dependent on technologies developed abroad. As a consequence, their functions are hidden in a black box, such as unshared source codes and online-limited distributions. Most commercial software programs are provided in combination with foreign software.

Table 1 shows a list of well-established genome sequencing software programs that appear in standard American college textbooks. As can be seen, software developed outside the U.S. are extremely rare. Although there are some excellent domestic application software such as those for protein structure prediction and gene annotation, basic software with widespread use as those listed in Table 1 are rarely found. The following are the possible reasons mentioned by some researchers in their interviews.

- Japan is far behind the U.S. in terms of the number of bioinformatics researchers. In the U.S., researchers can promptly switch their careers along with the shift in research trends. Such mobility in human resource has enabled quick securement of sufficient researchers specialized in bioinformatics. Human resource development in the bioinformatics area is an urgent issue for Japan.
- The integration of informatics with bioscience has not quite proceeded. The developer of BLAST was originally specialized in mathematics. We need an environment where researchers from different areas can work together.
- The development of software necessary for bioinformatics is abandoned before reaching its distribution. Working programs within individual research activities are left without further development. Even when a highly original algorithm is developed, the work is abandoned after its publication and does not lead to software development. Furthermore, the distribution of software requires packaging of manuals, installation tools and distribution media, which cannot be afforded by individual researchers.

An effective solution to these problems is the establishment of an environment for developing human resources specialized in bioinformatics or a system for evaluating the applicability of software and distributing them. A structure for providing governmental support to such systems needs to be discussed.

Table 1: Major genome sequencing software

Software	Inventor/creator	Characteristics
Homology search		
FASTA	Pearson 1988 (University of Virginia, U.S.)	Higher detection sensitivity than BLAST.
BLAST	Altschul 1990 (U.S. MCBI)	Higher speed than FASTA. Most commonly used.
PSI-BLAST	Altschul 1997 (U.S.)	Dialogic version of BLAST for searching protein families. Higher detection sensitivity than SSEARCH.
SEG	Wootton, Federhen 1993 (U.S.)	Increases comparison precision by excluding low-complexities and repeats.
SSEARCH	Pearson 1991 (U.S.)	Provides optimal alignment by using dynamic programming. Extremely slow.
Bayes block aligner	Zhu 1998 (U.S.)	Employs Bayes statistics. Slower than SSEARCH, but detects distantly related sequences.
PROBE	Neuwald 1997 (U.S.)	Similar function as PSI-BLAST. Finds most significant sequence set via non-dialogic process using Bayes statistics.
Multiple alignment (alignment of multiple sequences)		
ClustalW	Higgins, Sharp 1988 (U.K.)	Provides alignment of multiple sequences using a progressive approach. Most commonly used for multiple alignment.
PILEUP	Fen. Doolittle 1987 (U.S.)	Provides alignment of multiple sequences using a progressive approach. Employs the Needleman-Wunsch approach for sequence comparison.
MSA	Lipman 1989 (U.S.)	Provides optimal alignment via multidimensional dynamic programming.
PRRP	Goto 1996 (Japan National Institute of Advanced Industrial Science and Technology CBRC)	Constructs dendrograms and improves alignment via iterative learning.
SAGA	Notredame, Higgins (France)	Selects highly scored alignments using a genetic algorithm.
HMMER	Eddy 1998 (U.S.)	Employs the Hidden Markov model.
Profile search (search for characteristic patterns)		
ProfileSearch	Gribskov 1996 (U.S.)	Searches sequence patterns (motifs).
MAST	Bailey, Gribskov 1997 (U.S.)	Searches sequences matching gap-free sequence blocks.
Gene discovery		
RepeatMasker	Smit (University of Washington, U.S.)	Detects and removes repeats to facilitate gene discovery.
TWINSCAN	Korf (University of Washington, U.S.)	Compares genomes between different species and finds genes from conserved sequence regions. A hybrid of the alignment approach and ab initio approach.

Source: Authors' compilation based on reference [1]

4.4 Domestic projects in bioinformatics

While slowness in the progress of domestic bioinformatics research has been pointed out, there are some domestic research projects that can match U.S. and European researches.

4.4.1 Protein structure prediction program

With drug discovery in view, protein structure prediction is currently attracting great interest. In 2000, a team led by Professor Umeyama of Kitasato University developed FAMS and achieved high marks in CASP (the Critical Assessment of Techniques for Protein Structure Prediction), an international contest of protein structure prediction. FAMS outscored the programs developed in the U.S. and other countries that are commonly used in this area. Unlike other programs, which employ a bottom-up approach to construct an entire structure from partial structures, FAMS broadly grasps the entire structure before predicting the partial structures. This is how humans recognize structures, first focusing on the whole, and FAMS incorporated this human's view of structure recognition in its algorithm. Participation in such international

contests should promote domestic research and development.

4.4.2 Biogrid

Grid is an information technology that has gained great interest in recent years. It is an electric power-transmission network developed based on an image of "a computer system that allows you to use calculation power or discs freely just by plugging in, similar to electricity" (cited from "Trends in Grid Technology" in the February 2003 issue of Science and Technology Trends Quarterly Review).

Biogrid aims at the sharing of the program path necessary for bioinformatics through grid technology.

OBIGrid (Open Bioinformatics Grid) is led by "Genome Information Science," supported by Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology, and Initiative for Parallel Bioinformatics (IPAB). It aims at establishing an environment that enables the use of the latest databases and applications by just accessing the grid. OBIGrid has the potential for being a hub of genome analysis, lowering the barriers for new researchers entering the bioinformatics area. It should attract many researchers who, by themselves, cannot afford to arrange an environment for using various databases and applications required for bioinformatics research. Furthermore, the grid may provide the opportunity for disclosing experimental data that are usually hoarded. Since files can be easily accessed as in a LAN environment, novel discoveries from such experimental data can be expected.

4.4.3 Commercialization by business-academia collaboration

There is a move afoot to commercialize and distribute the software developed in public-funded bioinformatics research. The cDNA function annotation system developed in the FANTOM (Functional Annotation of Mouse) project of RIKEN was commercialized in 2002.The system is noteworthy as a successful case where the achievement of a cooperative research project between a public research institute and a private corporation led to a general-purpose package.

The cases mentioned above represent the potential of domestic bioinformatics research in Japan.

4.5 Efforts in human resource development

4.5.1 Talents sought for bioinformatics research

When researchers were interviewed about the causes of the slowness in domestic bioinformatics research, the predominant answer was "the lack of human resource." Bioinformatics is essential as a technology that strongly promotes diverse analyses ranging from genome and DNA analyses to structural and functional analyses of proteins. Since bioinformatics is an amalgam of informatics and bioscience, human resource development is an important task in our country. Bilingual talents understanding the languages of both biology/medicine and informatics are desired. Furthermore, the lack of communication and transaction across divisions and departments was also blamed for the slowness in domestic informatics research. In the FANTOM project of RIKEN, researchers from bioscience, medicine and informatics worked in close cooperation, which led to a great achievement. It is difficult to exploit mutual research results between informatics and bioscience areas merely by exchanging experimental and analytical data. The researchers in these two areas must not draw a distinction between their roles in such cooperative projects.

Regarding informatics specialists, those skilled in DB structure and programming are preferred for dealing with the enormous amount of data. Meanwhile, bioscientists are not asked to have great knowledge in IT, but are required to have sufficient understanding in the mechanisms of analytical programs and good command of them. Whenever needed, they should be capable of modifying the programs according to the individual experiments.

An effective way to develop human resources seems to be the promotion of human resource flow from the IT area with abundant talents into the bioscience area. However, a high level of bioinformatics research cannot be achieved

without a deep understanding of bioscience and medicine. Therefore, we must create a path for bioscience/medicine specialists to study IT and proceed into the bioinformatics area to secure high quality human resources.

In Western countries, especially in the U.S., researchers who have mastered informatics find their way into new research areas such as genome science. Their collaboration with genetics researchers led to advances in the bioinformatics area. Theoretically and mathematically supported analytical approaches have been applied to gene expression or protein structure/function experiments, and after repeating trials and errors, they were finally established as practical analytical techniques.

4.5.2 Policy for human resource development

Bioinformatics, moreover, is greatly expected as a new academic area. Due to the radical advances in genome science and protein research, an immediate supply of human resource is demanded by many research institutes including private institutes. Human resource development in the bioinformatics area is an urgent task, which should be promoted by the establishment of an environment having the following conditions.

(1) Graduate schools should offer students the option of studying in both areas of informatics and bioscience/medicine, and approve research in the amalgam area.
(2) An environment where informatics researchers or technicians can collaborate with bioscience/medical researchers must be established. This should promote the integration of experiments with bioinformatics. For example, analytical algorithms may be developed according to the progress of DNA or protein experiments, leaving the actual analysis to computers. This should promote technical advancement and development of practical talents.
(3) A system for evaluating the achievements and technical contributions of inventors of various computer analysis algorithms and software tools used in bioinformatics should be established.

For developing human resources for informatics, flexible management or modification of graduate school curricula should be effective. Additionally, faculty positions must be established in the bioinformatics area to offer bioinformatics researchers a career path comparable to those in existing areas. Evaluation of research achievements and human resource allocation must be performed within such framework. Furthermore, development of research systems accepting the participation of private institutes is desired. Such policies can also attract new entrants from the related academic areas. Especially, research areas such as mathematics, statistics and mathematical engineering can greatly contribute to the development of the theoretical foundation in bioinformatics. Governmental funding should be provided for systems meeting these requirements.

In 2001, Keio University established Advanced Life Science Institute Inc. in Tsuruoka City, Yamagata prefecture. With the slogan of "IT-driven bioscience," the institute provides an environment where faculties and young researchers in informatics and bioscience areas, together with the students, can study the mutual areas and conduct amalgam research. In 2002, Osaka University established the Graduate School of Frontier Biosciences, which is an interdisciplinary department composed of life science-related laboratories from Osaka University, including laboratories of medical/bioscience, bioengineering, biology and physics.

Meanwhile, in 2001, the government established human resource education units, supported by Special Coordination Funds for Promoting Science and Technology, for the prompt education of bioinformatics professionals. By 2002, a total of 6 units were established in Tokyo University, Kyoto University, the National Institute of Advanced Industrial Science and Technology, Keio University and the Nara Institute of Science and Technology, and integrated human resource development is progressing in each unit. By spreading such movements to other universities and research institutes, activation of bioinformatics and other interdisciplinary research areas should be promoted.

4.6 Conclusion

Speed is an important factor in genome research, the area in which bioinformatics exerts its strength. Many countries are making rapid investments of research resources in this area, and Japan is also increasing its investment in genome research. Yet, the development of domestic human resources in bioinformatics, which support genome research accelerating toward drug discovery and tailor medicines, has not reached an adequate level.

Genome and protein research is gradually shifting into a new stage, from relatively simple sequencing into functional analysis and, furthermore, into application. Along with this trend, requirements for bioinformatics will grow larger, as well as its importance in research studies. To fulfill such requirements, it is urgent that talents having academic/technical knowledge of both informatics and bioscience are developed. As mentioned in Chapter 4.5, some actions have already been taken, but not quite enough on a national scale.

For human resource education, in the short term, it should be effective to establish an environment where researchers from informatics and bioscience/medical areas can conduct their research while mutually sharing their knowledge and know-how to promote collaboration between informatics researchers/technicians and bioscience/medical researchers. Bioinformatics-related projects currently in progress should also put emphasis on this point. Meanwhile, we must promote mutual exchange of researchers and technicians between informatics and bioscience/medicine.

Acknowledgement

We would like to thank Project Director Dr. Yoshihide Hayashizaki, Team Leader Dr. Yasushi Okazaki and Project Director Dr. Akihiko Konagaya of RIKEN Genomic Sciences Center for kindly providing us with their information and resources in preparing this manuscript.

References

[1] Written by David W. Mount, translation supervised by Yasushi Okazaki and Hidemasa Bono, "Bioinformatics," Medical Science International (2002).

(Original Japanese version: published in January 2003)